Data Challenge Report

by Olena Verbytska

November 12, 2023

Data challenge: https://challengedata.ens.fr/participants/challenges/89/

1 Data preprocessing

1.1 Labeling

Variable **'date'** was treated as it has data type *datetime64*. Then I extracted encoded information about the day of the week and saved it in the variable **'day'**.

Variable **'station'** was transformed by *sklearn.preprocessing.LabelEncoder()*.

1.2 Imputations

Lag variables (**'p_q_'**) are missing, if the lag does not exist for the stop because of timetable structure. Therefore, it was reasonable to fill missing values with zeros.

To treat missing values of variable **'hour'** I labeled them as -1 and used *sklearn.preprocessing.LabelEncoder()* for all non-missing hours. (Before I tried to predict missing hours using two models: random forest classifier and multinomial logistic regression with explanatory variables **'station'** and **'train'**. The both scores for validation samples were approximately 0.98, which was not as good as I expected.)

1.3 Dropped variables

Variables **'way'** and **'composition'** did not take part in model training, because they basically take only one unique value.

2 Model selection

I used *lazypredict.Supervised.LazyRegressor()*, which applies more than 40 basic models to the given data and outputs a table of models' names and the corresponding scores. The model which have the smallest mean absolute error is *sklearn.ensemble.ExtraTreesRegressor()*

3 Hyperparameter optimization

For optimization of hyperparameters such as **n_estimators** and **min_samples_split** I used framework *optuna*. It is faster then using *sklearn.model_selection.GridSearchCV()* and its results are better optimized than the results of *sklearn.model_selection.RandomizedSearchCV()* within the same amount of trials.

4 The final model

sklearn.ensemble.ExtraTreesRegressor(n_estimators=3000, min_samples_split=8)
The public score is 0.0102552804598608 (Nov. 12, 2023, 6:18 p.m.).
The private score is 0.0109474271001274 (Oct. 18, 2023, 10:02 p.m.).

5 Epilogue

As may be expected, nearly 95% of my work is not included in the report. I did quite deep study of data even though the final data preprocessing is so simple. I am aware of data leak in this challenge, but I have not worked on it yet, as I wanted to spend more time on machine learning methods and optimization.

It was my first data challenge and, for sure, it will not be the last. I really enjoyed learning a lot of new things. Thank you for introducing such an interesting activity!